DOCUMENT RESUME

ED 466 638 TM 034 248

AUTHOR	Schnipke, Deborah L.
TITLE	The Accuracy of Pass/Fail Decisions in Random and Difficulty-Balanced Domain-Sampling Tests.
PUB DATE	2002-04-02
NOTE	10p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002). Some charts may not reproduce well.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Decision Making; Difficulty Level; *Item Banks; *Pass Fail Grading; Simulation; *Test Construction; Test Items
IDENTIFIERS	*Accuracy; Domain Knowledge; Randomization

ABSTRACT

A common practice in some certification fields (e.g., information technology) is to draw items from an item pool randomly and apply a common passing score, regardless of the items administered. Because these tests are commonly used, it is important to determine how accurate the pass/fail decisions are for such tests and whether fairly small, simple changes can be made to such tests to improve their psychometric properties without being too large a burden on the testing program. This simulation study compared a random test with a difficulty-balanced version of the test using 4 tests simulated for each of 1,000 examinees. Ability estimates and pass/fail decisions were slightly less accurate for the random test than for the difficulty-balanced test. The difference in difficulty distributions between the two test designs was dramatic. In addition, test takers who ended up failing the random test had harder tests on average than test takers in the difficulty-balanced test design, and those who ended up passing the random test had easier tests on average than test takers in the difficulty-balanced test design. From a fairness (and legal defensibility) point of view, the random test design is very undesirable. However, balancing tests on item difficulty is relatively easy to do. And this produced tests that are of equal difficulty across test takers and ability estimates and pass/fail decisions that are more accurate. Based on these results, it is recommended that testing programs that currently use random test designs switch to a difficulty-balanced test design. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 466 638

The Accuracy of Pass/Fail Decisions in Random and Difficulty-Balanced Domain-Sampling Test

Deborah L. Schnipke

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D. Schnipke

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034248

Paper Presented at the:
Annual Meeting of the National Council of Measurement in Education
New Orleans, LA - April 2002

BEST COPY AVAILABLE

The Accuracy of Pass/Fail Decisions in Random and Difficulty-Balanced Domain-Sampling Tests[1]

Deborah L. Schnipke
CAT*ASI

Abstract. A common practice in some certification fields (e.g., information technology) is to randomly draw items from an item pool and apply a common passing score, regardless of which items were administered. Because these tests are commonly used, it is important to determine how accurate the pass/fail decisions are for such tests and whether fairly small, simple changes can be made to such tests to improve their psychometric qualities without being too large of a burden on the testing programs. The purpose of this project was to compare a random test with a difficulty-balanced version of the test. Ability estimates and pass/fail decisions were slightly less accurate for the random test than for the difficulty-balanced test. The difference in difficulty distributions between the two designs was dramatic. In addition, test takers who ended up failing the random test had harder tests on average than test takers in the difficulty-balanced test design, and those who ended up passing the random test had easier tests on average than test takers in the difficulty-balanced test design. From a fairness (and legal defensibility) point of view, the random test design is very undesirable. However, balancing tests on item difficulty is relatively easy to do, and this produces tests that are of equal difficulty across test takers and ability estimates and pass/fail decisions that are more accurate. Based on these results, it is recommended that testing programs that currently use random test designs switch to a difficulty-balanced test design.

Introduction

A common practice in some certification fields (e.g., information technology) is to use a domain sampling approach to testing. The domain for the test is defined, and a standard for being certified is determined based on the percentage of the domain that needs to be answered correctly. Items are randomly drawn from the domain (with or without content balancing), and the resulting test forms are often not equated in any way. A common passing score (in terms of number or percentage of items answered correctly) is used, regardless of which items were administered.

The purpose of this project is to compare the random (unequated) domain-sampling test with a version that takes the statistical properties of the items into account. An adaptive version of the test will also be compared. The goal is to determine the return on investment for adding various levels of complexity to the testing process.

Methods

A two-part process was followed for the simulations. The first part emulates standard practice in IT certification for gathering pretest data – the beta test. In

[1] Presented at the Annual Meeting of the National Council on Measurement in Education, April 2002, New Orleans.

the beta test, the entire item pool was administered to a small group of simulated test takers to gather data to calculate preliminary statistics for determining which items keep in the final pool. In the second part, a large number of simulated test takers took 4 different tests based on the final item pool. The 4 tests used different test designs: a random test, a difficulty-balanced test, an adaptive test, and a test using the whole item pool. The item parameters, ability distributions, beta test simulations, and the main simulations are described in more detail in the following sections.

Item Parameters

The three-parameter logistic (3PL) model was used to simulate "noise" in the data. The original item pool had 150 simulated items. The difficulty parameter was drawn from a standard normal distribution. The discrimination parameter was drawn from a normal distribution with a mean of .80 and standard deviation of .20, with a correlation of 0.20 built into the relationship between difficulty and discrimination. The lower asymptote parameter was drawn from a uniform distribution ranging from 0 to 0.30. The distributions for the item parameters were based on results from an operational testing program.

Beta Test

A preliminary beta test simulation was run with 100 test takers (with standard normal ability estimates) responding to all 150 items so that classical item statistics (proportion correct and item-test correlation) and Rasch item difficulty and infit and outfit statistics could be calculated. These statistics will be referred to as the beta statistics. The beta test simulation emulates standard practice in IT certification for gathering pretest data.

The beta statistics were used to flag items with unacceptable statistical characteristics. Using liberal flagging criteria, items were flagged if

- the p value (proportion correct) was less than .1 or greater than .95,
- the item-test correlation (item discrimination) was not significantly greater than 0 (at α=.10), or
- the Rasch infit or outfit statistic[2] was greater than 1.5.

Based on these flagging rules, 9 items were removed (8 because of non-significant item-test correlations and 1 because of large outfit). Therefore, the final item pool contained 141 items.

[2] The infit and outfit statistics are based on the residual analysis of the difference between the expected score and the observed score. The infit statistic is weighted by the ability distribution, whereas the outfit statistic is not weighted (and is therefore more affected by outliers). When the infit or outfit is standardized it can be interpreted as a z score. Items having infit or outfit greater than 1.5 have a difference between the observed score and the expected score that exceeds one and a half standard errors.

Main Simulation

To simulate examinees, 1000 ability values were drawn from a standard normal distribution. Four tests were simulated for each of these 1000 examinees:

- **Whole Pool Test**. A test containing all 141 items in the final pool. The true 3PL parameters were used for determining the response (right/wrong).
- **Random Test**. A test of 50 items (using the true 3PL parameters for determining the response) where the items were selected completely at random.
- **Difficulty Balanced Test**. A test of 50 items (using the true 3PL parameters for determining the response) that has the same number of items from each classical difficulty bin (see Table 1) for each examinee so that the distribution of difficulty for the test is controlled.
- **Adaptive Test**. An adaptive test of 50 items (using the true 3PL parameters for determining the response). Items were selected based on maximum information using the 3PL parameters. Ability estimates during the simulation were calculated via Bayes modal scoring with a standard normal prior.

Table 1. Difficulty bins for the difficulty-balanced test. Bins were based on the p values from the beta data. During test administration, 3 items were administered from Bin 1, 5 items from Bin 2, etc.

Difficulty Bins	Min p Value	Max p Value	Number in Item Pool	Number on Test
1	.80	.95	10	3
2	.70	.79	14	5
3	.66	.69	13	5
4	.60	.65	18	6
5	.55	.59	12	4
6	.50	.54	24	9
7	.45	.49	14	5
8	.40	.44	13	5
9	.30	.39	18	6
10	.13	.29	5	2
Total			141	50

Results

The return on investment for adding levels of complexity to the testing process was investigated by comparing

- the precision of the ability estimates for each of the tests,
- the accuracy of the pass/fail decisions for each of the tests, and
- the difficulty of the items administered to each test taker

Precision of Ability Estimates

Table 2 shows the correlations between simulated true ability and estimated ability from each of the four tests. The results are in the expected order. The whole pool test was most precise, followed by the adaptive test, difficulty-

balanced test, then lastly the random test. Figure 1 shows these results graphically. Note that even the difficulty-balanced and random tests recovered the true abilities fairly well.

Table 2. Correlations between true ability and estimated ability.

	Pearson Correlation	Sig. (2-tailed)	N
True Ability	1	.	1000
Estimated Ability: Whole Pool	.982**	.000	1000
Estimated Ability: Random	.943**	.000	1000
Estimated Ability: Difficulty Balanced	.950**	.000	1000
Estimated Ability: Adaptive	.969**	.000	1000

**. Correlation is significant at the 0.01 level (2-tailed).

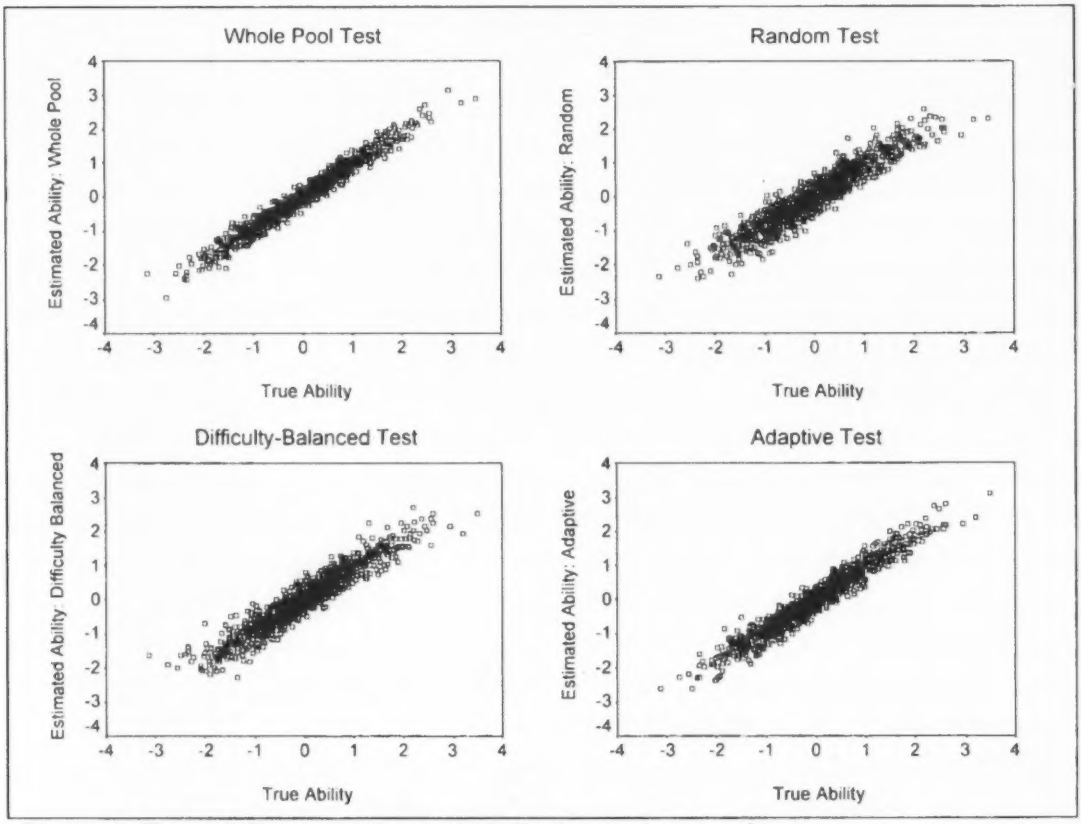

Figure 1. True ability vs. estimated ability for each type of test.

Accuracy of Pass/Fail Decisions for Each Test

As is done for tests in some fields, the pass/fail cutoff was based on the percentage of questions answered correctly, regardless of which items were

administered to the test taker.[3] For this study, 80% correct was used. Decisions based on the whole pool were used as the standard against which to compare the decisions for the random and difficulty-balanced tests.

As shown in Table 3, the decisions for the difficulty-balanced test are more accurate than the decisions for the random test. For the difficulty-balanced test, there were 934 correct decisions (out of 1000): 120 passed on the difficulty-balanced test and the Whole Pool test, and 814 failed both tests. For the random test, there were 926 correct decisions (out of 1000): 115 passed on the random test and the Whole Pool test, and 811 failed both tests.

Table 3. Pass/fail decisions for the whole pool test vs. the random and difficulty-balanced tests.

		Pass/Fail Status For Whole Pool Test		Total
		Fail	Pass	
Total		857	143	1000
Pass/Fail Status For Random Test	Fail	**811**	28	839
	Pass	46	**115**	161
Pass/Fail Status For Difficulty-Balanced Test	Fail	**814**	23	837
	Pass	43	**120**	163

Difficulty of Items Administered to Each Test Taker

The final comparison between types of tests was based on the difficulty of items administered to each test taker. For the random and difficulty-balanced tests, the mean, standard deviation, minimum, and maximum p values (proportion correct) were calculated for each test taker (displayed in the rows of Table 4), and the minimum, maximum, mean, and standard deviation were calculated across test takers (displayed in the columns of Table 4) for each of the within test taker statistics. For example, for the random test, the mean of the mean p values was .5510, and the standard deviation of the mean p values was .01756. Note that the standard deviation of the mean p values for the difficulty-balanced test (.00301) was much smaller than the standard deviation for the random test (.01756).

Table 4. Summary of p values for random and difficulty-balanced tests.

		N	Minimum	Maximum	Mean	Std. Deviation
Random Test	Mean(p_value)	1000	.49	.61	.5510	.01756
	SD(p_value)	1000	.12	.19	.1562	.01144
	Min(p_value)	1000	.13	.35	.2163	.06780
	Max(p_value)	1000	.75	.94	.9059	.04339
Difficulty Balanced Test	Mean(p_value)	1000	.54	.56	.5529	.00301
	SD(p_value)	1000	.15	.18	.1609	.00509
	Min(p_value)	1000	.13	.28	.1981	.06000
	Max(p_value)	1000	.82	.94	.9126	.03569

[3] Because percentage correct scores on an adaptive test are meaningless adaptive test results are not included in this section. There is no direct relationship between the cutoffs based on number right scores and 3PL IRT ability estimates.

The difference between mean p values for the random and difficulty-balanced tests is more obvious when displayed graphically, as shown in Figure 2. The distribution of mean p values is clearly much more spread out for the random test than for the difficulty-balanced test, as shown in Figure 2. Thus the items that the test takers received were more variable in difficulty for the random test design. In the random test design, some test takers were lucky and received an easier test (higher mean p value), whereas other test takers were unlucky and received a more difficult test (lower mean p value). In the difficulty-balanced test design, the difficulty across test takers was very similar across test takers. (The same scales are used for the axes in both graphs in Figure 2.)

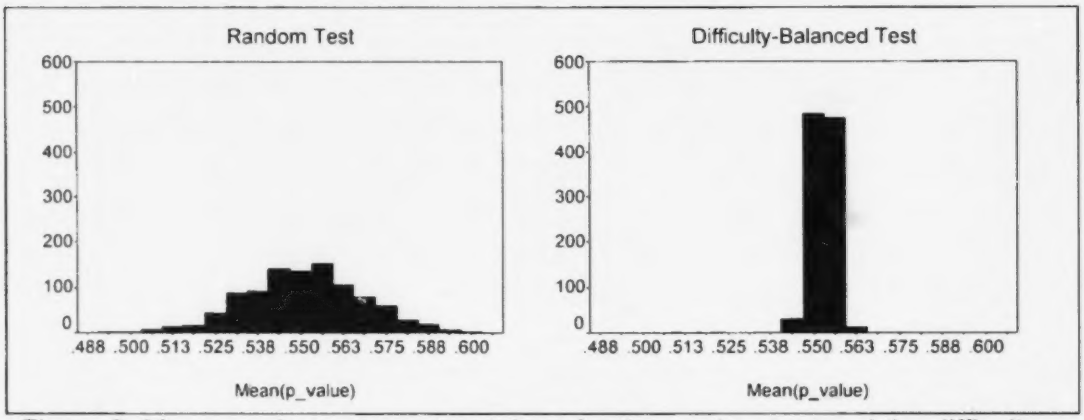

Figure 2. Mean p values across test takers for the random test and the difficulty-balanced test.

As shown in Table 5 and Figure 3, the mean p value varied by pass/fail status for the random test, but not for the difficulty-balanced test. Test takers who ended up failing the random test had a harder test on average (.5506) than those who ended up passing the random test (.5533).

In addition, test takers who ended up failing the random test had a harder test on average than the average test in the difficulty-balanced test design (.5529), and those who ended up passing the random test had an easier test on average than the average test in the difficulty-balanced test design.

Table 5. Summary of p values for random and difficulty-balanced tests by pass/fail status.

Test	Status		N	Minimum	Maximum	Mean	Std. Deviation
Random Test	Fail	Mean(p_value)	839	.49	.61	.5506	.01764
		SD(p_value)	839	.12	.19	.1564	.01143
		Min(p_value)	839	.13	.34	.2141	.06835
		Max(p_value)	839	.75	.94	.9060	.04387
		Valid N (listwise)	839				
	Pass	Mean(p_value)	161	.51	.60	.5533	.01704
		SD(p_value)	161	.12	.19	.1548	.01139
		Min(p_value)	161	.13	.35	.2279	.06381
		Max(p_value)	161	.80	.94	.9053	.04090
		Valid N (listwise)	161				
Difficulty Balanced Test	Fail	Mean(p_value)	837	.54	.56	.5529	.00297
		SD(p_value)	837	.15	.18	.1609	.00512
		Min(p_value)	837	.13	.28	.1979	.06020
		Max(p_value)	837	.82	.94	.9124	.03571
		Valid N (listwise)	837				
	Pass	Mean(p_value)	163	.54	.56	.5529	.00322
		SD(p_value)	163	.15	.17	.1611	.00493
		Min(p_value)	163	.13	.28	.1990	.05914
		Max(p_value)	163	.82	.94	.9133	.03570
		Valid N (listwise)	163				

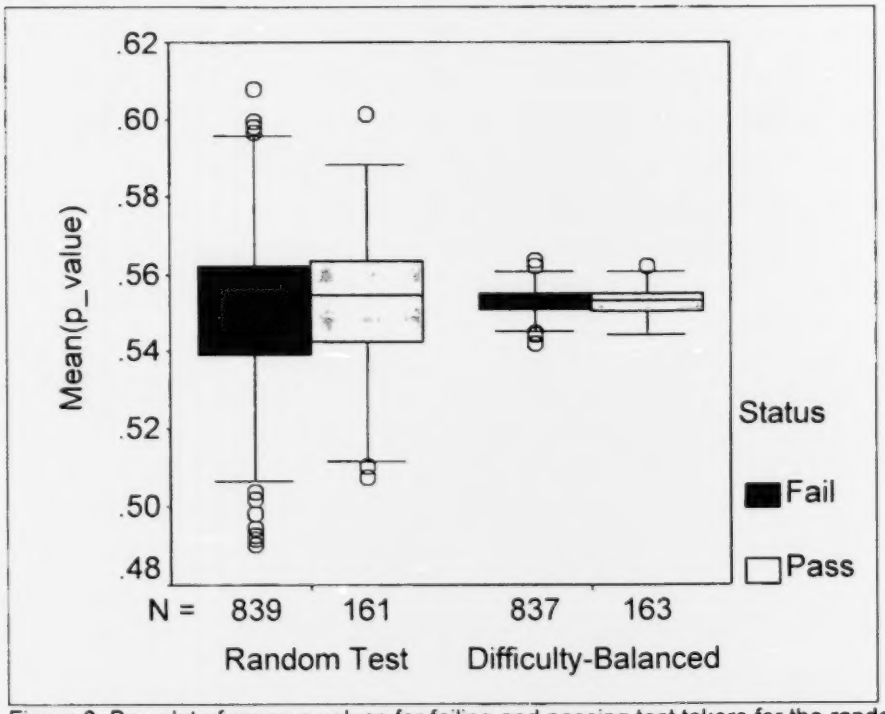

Figure 3. Box plot of mean p values for failing and passing test takers for the random and difficulty-balanced tests. Lower values of Mean(p_value) indicate more difficult tests on average.

BEST COPY AVAILABLE

Discussion

Psychometricians generally do not recommend random, unequated tests for obvious reasons (e.g., fairness and score comparability). Such tests are commonly used, however, in some fields. It is important to determine how accurate the pass/fail decisions are for such tests and whether fairly small, simple changes can be made to such tests to improve their psychometric qualities without being too large of a burden on the testing programs.

In the present study, the primary comparison was between a random test and a test that was balanced on difficulty. The difficulty-balanced test was constructed by creating 10 difficulty bins based on p values from the beta-testing (field-testing) phase. For the difficulty-balanced test, the number of items selected from each difficulty bin for each test taker during test administration matched the proportion of items in those bins in the whole pool.

Ability estimates and pass/fail decisions were slightly less accurate for the random test than for the difficulty-balanced test. Not surprisingly, the average difficulty of the random tests was more variable than the average difficulty for the difficulty-balanced tests. The difference between the two designs was dramatic. In addition, test takers who ended up failing the random test had harder tests on average than test takers in the difficulty-balanced test design, and those who ended up passing the random test had easier tests on average than test takers in the difficulty-balanced test design.

From a fairness (and legal defensibility) point of view, the random test design is very undesirable. However, balancing tests on item difficulty is relatively easy to do, and this produces tests that are of equal difficulty across test takers and ability estimates and pass/fail decisions that are more accurate. Based on these results, it is recommended that testing programs that currently use random test designs switch to a difficulty-balanced test design.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM034248

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Accuracy of Pass/Fail Decisions in Random and Difficulty-Balanced Domain-Sampling Tests	
Author(s): Deborah L Schnipke	
Corporate Source: CAT@ASI	Publication Date: 4/2/2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: Deborah L Schnipke / Research + Development Lead
Organization/Address: CAT@ASI, 1007 Church St, 7th Floor, Evanston, IL 60201	Telephone: 847-866-2146 FAX: 860-371-3209
	E-Mail Address: dschnipke@gatinc.com Date: 4/30/2002

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)